Module 2: Body Fat Prediction Model

Group 13: Ming Pei, Shravan Kaul, Zheming Cao

1 Introduction

Here we have a dataset includes body fat and various body circumferences measurements of 252 men. We aim to build a simple but robust model to predict body fat effectively. By cleaning data, stepwise feature selection using BIC, we choose a linear model that match our expectation.

2 Data Cleaning

We find that body fat is highly correlated to density, adiposity, chest and abdomen. Moreover, most of the correlation coefficients between weight, adiposity and body circumferences are greater than 0.7, which means they are highly correlated.

In data cleaning part, we impute 2 values and remove 3 outliers. Using interquartile range (IQR) method, we detect 5 outliers at first. Then we calculate adiposity using weight and height, as for body fat, using density[1], and compare them with corresponding values in dataset. Finally, we impute height of the 42^{nd} man using his weight and adiposity and body fat of the 182^{nd} man by the value obtained from an online body fat calculator[2]. We remove the other 3 outliers and fit models with the remaining data of 249 men.

3 Model

Our final model is

$$BodyFat\% = 0.881 \times Abdomen(cm) - 0.083 \times Weight(lbs) - 1.357 \times Wrist(cm) - 22.942.$$

The predictors are abdomen(cm), weight(lbs) and wrist(cm). For a man with abdomen circumference 85 cm, weight 155 lbs and wrist circumference 18 cm, his body fat should be 14.6%. The 50% prediction interval of body fat for the man is between 11.97% and 17.32%. The estimated coefficient of abdomen is 0.881, which means an unit increase in abdomen circumference keeping the measurement of weight and wrist unchanged will lead to a 0.881% increase in body fat. The other two coefficients can be interpreted similarly.

3.1 Motivation and Model Selection

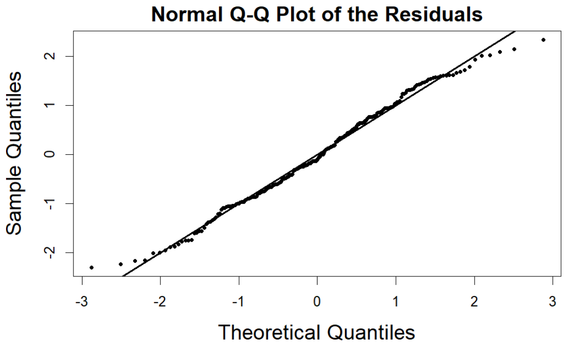
Our aim is to find a simple but robust model, hence we decided to fit a linear model to predict body fat. We perform stepwise variable selection with BIC, because BIC penalizes the model more for its complexity compared with AIC. The stepwise selection is performed in forward, backward and both directions and gives us three models in table 1. Besides BIC, we also use

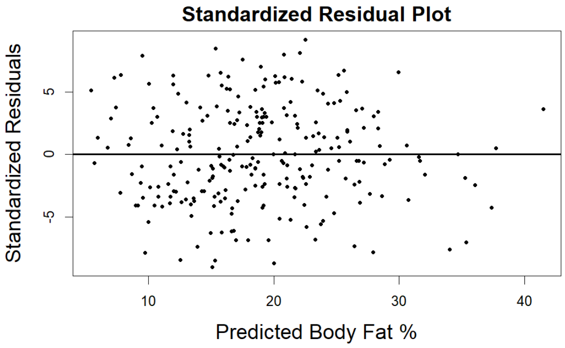
adjusted R^2 to compare the goodness of fit of three models.
The forward selection model has the lowest BIC score and the highest adjusted R^2, which means it has the best fit among three models. Since weight is highly correlated with both abdomen and wrist, we fitted models with an interaction term of abdomen and weight and another one with an interaction of wrist and weight. The interaction terms in two models are not significant at a significance level of 0.05, thus the full forward selection model is better.

Direction	Predictors	BIC	Adjusted R^2
Forward	Abdomen, weight, wrist	1414.1	0.7342
Backward	Age, abdomen, wrist	1417.8	0.7302
Both	Abdomen, wrist, height	1415.4	0.7328

Table 1: Model Comparison

3.2 Model Diagnostics





We assume the effects are linear and errors follow normal distribution. By observing the Q-Q plot, we see that residuals follow normality for the majority of distribution. When residual becomes larger, points deviates to right side, smaller residuals points deviates to left and hence tails are created. It means points with larger residual deviates more than expected. That means our model overestimate people with lower body fat and higher body fat.
Residuals are randomly distributed in the standard residual plot, shows model follows independence assumption.

4 Suggested Improvements

We built a simple model with good precision. Improvements can be focused on fitting more bodyfat deviated population, a more complicated model could be taken into consideration like non linear or regression by steps if possible. We can also use machine learning models to achieve higher accuracy.

5 Conclusion

In all, we built a pretty simple model to predict bodyfat with clear linear effects. One can easily estimate bodyfat using our model which is easy to interpret and robust.

Contribution

Ming Pei: code for imputation, data cleaning and model building; Section 2 and Section 3 in report; Data Cleaning and Model Selection in slides.

Shravan Kaul: code for outlier detection, model building and shiny app; final model in Section 3 in report; Final Model Summary, Rule of Thumb and Conclusion in slides.

Zheming Cao: code of model diagnostics; Section 1, Section 3.2, Section 4 and 5 in report; Assumption, Strength and Weakness and Conclusion in slides.

References

[1] Katch, Frank and McArdle, William (1977). Nutrition, Weight Control, and Exercise, Houghton Mifflin Co., Boston.

[2] Webpage body fat calculator https://www.calculator.net/body-fat-calculator.html